

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-315207

(43)Date of publication of application : 14.11.2000

(51)Int.Cl. G06F 17/30
G06F 17/21

(21)Application number : 11-123487

(71)Applicant : JUST SYST CORP

(22)Date of filing : 30.04.1999

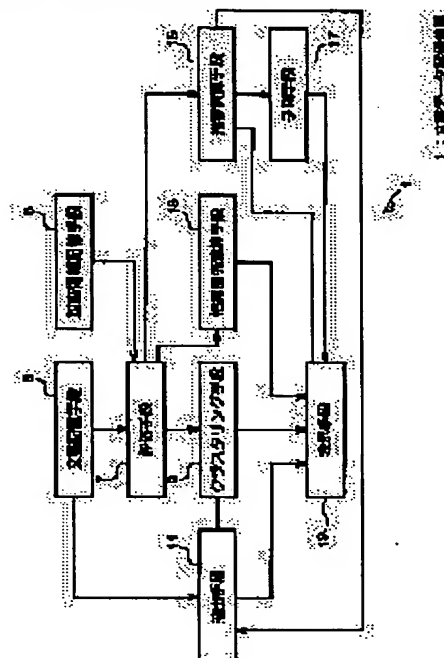
(72)Inventor : TAKATO ATSUSHI
MITOBE KATSUHIKO
DOI KATSUYUKI
MITSUYA HIROYUKI

(54) STORAGE MEDIUM IN WHICH PROGRAM TO EVALUATE DOCUMENT DATA IS STORED

(57)Abstract:

PROBLEM TO BE SOLVED: To obtain evaluation matched to an operator's intention for each document.

SOLUTION: Plural documents are stored in a document storage means 3. A viewpoint profile and viewpoint profile corresponding information to be composed of plural related words related to the viewpoint profile are stored in a corresponding information storage means 5. An evaluated value for each viewpoint profile is calculated and the evaluated value of each viewpoint profile is decided for each piece of document data to be evaluated by an evaluating means 7. Clustering is performed based on the evaluated value of each piece of document data by a clustering means 9. A characteristic keyword regarding a certain cluster is extracted from the document data sorted into the cluster by comparing pieces of document data sorted into different clusters by an extracting means 11. The extracted keyword is displayed on a display means 19.



LEGAL STATUS

[Date of request for examination] 16.02.2000

[Date of sending the examiner's decision of rejection] 07.07.2003

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

書誌

- (19)【発行国】日本国特許庁(JP)
(12)【公報種別】公開特許公報(A)
(11)【公開番号】特開2000-315207(P2000-315207A)
(43)【公開日】平成12年11月14日(2000. 11. 14)
(54)【発明の名称】文書データを評価するプログラムを記憶した記憶媒体
(51)【国際特許分類第7版】

G06F 17/30

17/21

【FI】

G06F 15/401 310 D

15/20 570 N

590 E

15/40 370 A

【審査請求】有

【請求項の数】17

【出願形態】OL

【全頁数】12

(21)【出願番号】特願平11-123487

(22)【出願日】平成11年4月30日(1999. 4. 30)

(71)【出願人】

【識別番号】390024350

【氏名又は名称】株式会社ジャストシステム

【住所又は居所】徳島県徳島市沖浜東3-46

(72)【発明者】

【氏名】高藤 淳

【住所又は居所】徳島県徳島市川内町平石若松108番4号 株式会社ジャストシステム内

(72)【発明者】

【氏名】水戸部 勝彦

【住所又は居所】徳島県徳島市川内町平石若松108番4号 株式会社ジャストシステム内

(72)【発明者】

【氏名】土居 功志

【住所又は居所】徳島県徳島市川内町平石若松108番4号 株式会社ジャストシステム内

(72)【発明者】

【氏名】三ツ矢 浩之

【住所又は居所】徳島県徳島市川内町平石若松108番4号 株式会社ジャストシステム内

(74)【代理人】

【識別番号】100092956

【弁理士】

【氏名又は名称】古谷 栄男（外2名）

【テーマコード(参考)】

5B009

5B075

【Fターム(参考)】

5B009 SA12 SA14 VA02

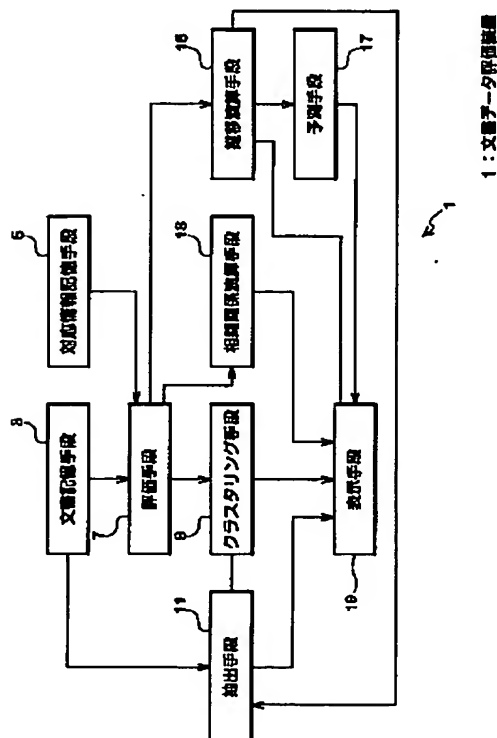
5B075 ND03 NK02 NR03 NR12 PP02 PP03 PP11 PP22 PR06 QM08 QS01 UU06

要約

(57)【要約】

【課題】各文書について、操作者の意図に合致した評価を得る。

【解決手段】文書記憶手段3は、複数の文書が記憶される。対応情報記憶手段5は、視点プロフィールとこの視点プロフィールに関連する複数の関連ワードで構成される視点プロフィール対応情報を複数記憶する。評価手段7は、前記評価対象の各文書データについて、前記各視点プロフィール毎の評価値を求めて、各視点プロフィールの評価値を決定する。クラスタリング手段9は、前記各文書データの評価値に基づいて、クラスタリングする。抽出手段11は、異なるクラスタに分類された文書データを比較して、あるクラスタに分類された文書データから、そのクラスタに関して特徴的なキーワードを抽出する。かかる抽出されたキーワードは表示手段19に表示される。



請求の範囲

【特許請求の範囲】

【請求項1】入力装置、制御装置、および記憶装置を備えたコンピュータを文書データ評価装置として機能させるプログラムを記憶した記録媒体において、前記プログラムは、前記コンピュータに、以下の処理を実行させること、視点プロファイルとこの視点プロファイルに関連する複数の関連ワードで構成される視点プロファイル対応情報を複数用いて、前記文書データについて、前記各視点プロファイル毎の評価値を求めて、各視点プロファイルと同じ次元を持つ数値ベクトルを当該文書データの評価値とする、を特徴とするプログラムを記憶した記録媒体。

【請求項2】請求項1のプログラムを記憶した記録媒体において、前記複数の視点プロファイルには、逆の方向性を有する1対の視点プロファイルを含んでいること、を特徴とするもの。

【請求項3】請求項2のプログラムを記憶した記録媒体において、前記視点プロファイルは、状態を表すプロファイルであること、を特徴とするもの。

【請求項4】請求項2のプログラムを記憶した記録媒体において、前記視点プロファイルは、変化を表すプロファイルであること、を特徴とするもの。

【請求項5】請求項1のプログラムを記憶した記録媒体において、前記評価対象の文書が複数あり、前記各文書データの評価値に基づいて、クラスタリングすること、を特徴とするもの。

【請求項6】請求項5のプログラムを記憶した記録媒体において、異なるクラスタにクラスタリングされた文書データを比較して、あるクラスタに分類された文書データから、そのクラスタに関して特徴的なキーワードを抽出すること、を特徴とするもの。

【請求項7】請求項1のプログラムを記憶した記録媒体において、前記評価対象の文書が複数あり、前記各文書データの評価値に基づいて、前記複数のプロファイルの相関関係を求めること、を特徴とするもの。

【請求項8】請求項1のプログラムを記憶した記録媒体において、前記評価対象の文書が複数あり、当該文書データを時系列に並べて、各視点プロファイルに関する前記各文書データの評価値の時間的変化を求めること、を特徴とするもの。

【請求項9】請求項1のプログラムを記憶した記録媒体において、前記評価対象の文書が複数あり、当該文書データを時系列に並べるとともに、前記任意複数の視点プロファイルに関して視点プロファイル毎に重み係数による総合評価値を求めて、前記総合評価値の時間的変化を求めること、を特徴とするもの。

【請求項10】請求項8または請求項9のプログラムを記憶した記録媒体において、前記求めた前記各文書データの評価値の時間的変化に基づいて、当該視点プロファイルにおける変化を予測すること、を特徴とするもの。

【請求項11】請求項8または請求項9のプログラムを記憶した記録媒体において、前記求めた各視点プロファイルに関する前記各文書データの評価値の時間的変化を報知すること、を特徴とするもの。

【請求項12】請求項8または請求項9のプログラムを記憶した記録媒体において、前

記求めた各視点プロファイルに関する前記各文書データの評価値の時間的変化に基づいて、変化の大きな時期の文書データを報知すること、を特徴とするもの。

【請求項13】請求項8または請求項9のプログラムを記憶した記録媒体において、前記求めた各視点プロファイルに関する前記各文書データの評価値の時間的変化に基づいて、変化の大きな時期の文書データを特定し、特定した文書データから、その文書に関して特徴的なキーワードを抽出すること、を特徴とするもの。

【請求項14】請求項1のプログラムを記憶した記録媒体において、前記評価対象の文書は複数あり、前記プログラムは、前記各文書からタームを抽出し、抽出したタームで各文書を数値化して記憶しておき、以下の1)～2)の処理を前記複数の視点プロファイルについて実行して、前記評価値を求めること、1)前記視点プロファイルの複数の関連ワードを数値化し、2)前記ターム数値化処理による値と前記視点プロファイル数値化処理による値に基づいて、当該視点プロファイルについての評価を演算する、を特徴とするもの。

【請求項15】請求項14のプログラムを記憶した記録媒体において、前記タームによる数値化処理では、前記各文書を構成する語句についてtfidf値を求めて、重要語句を複数決定し、各重要語句について前記各文書のtfidf値を演算して、前記複数の重要語句の数と同じ次元の数値ベクトルが求められ、前記各視点プロファイルによる評価値演算処理では、前記各視点プロファイルにおける関連ワードのtfidf値を演算して、前記関連ワードの数と同じ次元の数値ベクトルが各文書毎に求められ、前記両数値ベクトルの類似度が前記評価として演算されること、を特徴とするもの。

【請求項16】文書データを評価する文書データ評価方法であって、視点プロファイルとこの視点プロファイルに関連する複数の関連ワードで構成される視点プロファイル対応情報を複数記憶しておき、前記文書データについて、前記各視点プロファイル毎の評価値を求めて、各視点プロファイルと同じ次元を持つ数値ベクトルを当該文書データの評価値とすること、を特徴とする文書データ評価方法。

【請求項17】視点プロファイルとこの視点プロファイルに関連する複数の関連ワードで構成される視点プロファイル対応情報を複数記憶する対応情報記憶手段、評価対象の文書データについて、前記各視点プロファイル毎の評価値を求めて、各視点プロファイルと同じ次元を持つ数値ベクトルを当該文書データの評価値として決定する評価手段、を備えたことを特徴とする文書データ評価装置。

詳細な説明

【発明の詳細な説明】

【0001】

【発明の属する技術分野】この発明は、文書データを評価する文書データ評価装置に関し、特にプロファイルを用いた評価に関する。

【0002】

【従来技術およびその課題】今日、多くの文書情報をデータとして記憶しておき、これらの文書情報から所望の情報や知識を発見すること(以下、マイニングという)が試みられている。かかるマイニングをコンピュータでおこなうためには、その前提として各文書の内容を数値化する必要がある。

【0003】かかる文書の内容を正しく数値化する手法として、人間が文書を読んで内容を理解して、評価することも考えられる。しかし、評価する人間によって主観的な値となり、また、文書数が多いと複数人によって評価せざるを得ず、評価にばらつきが多いという問題がある。

【0004】かかる文書の数値化の手法として、従来、tfidf 値を用いた評価法が知られている。tfidf 値を用いた評価法とは、文書の要約処理等に用いられている手法である。

【0005】tfidf 値を用いた評価法について簡単に説明する。まず、記憶されている複数の文書から、出現頻度に基づいて数値化するための語句(以下タームという)を複数特定する。各文書について、各タームの出現頻度を求める。各文書について、タームの数と同じ次元のベクトル空間にてベクトル化した値(以下タームベクトルという)を求める。かかるタームベクトルをその文書の評価値とする(ベクトル空間モデルについては、岸田和明著「情報検索の理論と技術」(勁草書房)85頁～89頁に詳しい)を用いた。

【0006】しかし、このような評価方法では、より正確な評価をするために、タームの数を増やせば増やすほど、多次元の数値ベクトルとなり、前記数値ベクトルが示している文書全体の意味内容が把握できないという問題があった。

【0007】この発明は上記問題を解決し、各文書について、操作者の意図に合致した評価を得ることができる文書データ評価装置またはその方法を提供することを目的とする。また、複数の文書を操作者の検索意図に合致して数値化することができる文書データ評価装置またはその方法を提供することを目的とする。

【0008】

【課題を解決するための手段および発明の効果】1)本発明にかかるプログラムを記憶した記録媒体において、前記プログラムは、コンピュータに以下の処理を実行させる:視点プロフィールとこの視点プロフィールに関連する複数の関連ワードで構成される視点プロフィール対応情報を複数用いて、前記文書データについて、前記各視点プロフィール毎の評価値を求めて、各視点プロフィールと同じ次元を持つ数値ベクトルを当該文書データの評価値とする。これにより、前記文書データを客観的に、かつ正確に評価することができる。

【0009】2)本発明にかかるプログラムを記憶した記録媒体においては、前記複数の視点プロフィールには、逆の方向性を有する1対の視点プロフィールを含んでいる。このように逆の方向性を有する1対の視点プロフィールを用いて評価を行うことにより、より正確に評価することができる。

【0010】3)本発明にかかるプログラムを記憶した記録媒体においては、前記視点プロフィールは、状態を表すプロフィールである。したがって、前記文書データを当該状態に関して評価することができる。

【0011】4)本発明にかかるプログラムを記憶した記録媒体においては、前記視点プロフィールは、変化を表すプロフィールである。したがって、前記文書データを当該変化に関して評価することができる。

【0012】を特徴とするもの。

【0013】5)本発明にかかるプログラムを記憶した記録媒体においては、前記評価対象の文書が複数あり、前記各文書データの評価値に基づいて、クラスタリングする。

したがって、前記複数の文書データからより操作者の好む情報を取り出すことができる。

【0014】6)本発明にかかるプログラムを記憶した記録媒体においては、異なるクラスタにクラスタリングされた文書データを比較して、あるクラスタに分類された文書データから、そのクラスタに関して特徴的なキーワードを抽出する。これにより、クラスタリングされた文書データのより詳細な分析が可能となる。

【0015】7)本発明にかかるプログラムを記憶した記録媒体においては、前記評価対象の文書が複数あり、前記各文書データの評価値に基づいて、前記複数のプロフィールの相関関係を求める。これにより、前記複数のプロフィールの相関関係を分析することができる。

【0016】8)本発明にかかるプログラムを記憶した記録媒体においては、前記複数の文書データを時系列に並べて、各視点プロフィールに関する前記各文書データの評価値の時間的な変化を求める。これにより、注目する視点に関する時間的な履歴を得ることができる。

【0017】9)本発明にかかるプログラムを記憶した記録媒体においては、前記評価対象の文書が複数あり、当該文書データを時系列に並べるとともに、前記任意複数の視点プロフィールに関して視点プロフィール毎に重み係数による総合評価値を求めて、前記総合評価値の時間的な変化を求める。したがって、注目する複数の視点に関する総合評価の時間的な履歴を得ることができる。

【0018】10)本発明にかかるプログラムを記憶した記録媒体においては、前記求めた前記各文書データの評価値の時間的な変化に基づいて、当該視点プロフィールにおける変化を予測する。これにより、注目する視点に関する時間的な履歴を予想することができる。

【0019】11)本発明にかかるプログラムを記憶した記録媒体においては、前記求めた各視点プロフィールに関する前記各文書データの評価値の時間的な変化を報知する。これにより、注目する視点に関する時間的な履歴を操作者に報知することができる。

【0020】12)本発明にかかるプログラムを記憶した記録媒体においては、前記求めた各視点プロフィールに関する前記各文書データの評価値の時間的な変化に基づいて、変化の大きな時期の文書データを報知する。したがって、注目する視点に関して影響が大きいと思われる文書データを操作者に知らせることができる。

【0021】13)本発明にかかるプログラムを記憶した記録媒体においては、前記求めた各視点プロフィールに関する前記各文書データの評価値の時間的な変化に基づいて、変化の大きな時期の文書データを特定し、特定した文書データから、その文書に関して特徴的なキーワードを抽出する。これにより、注目する視点に関して影響が大きいと思われるキーワードを操作者に知らせることができる。

【0022】14)本発明にかかるプログラムを記憶した記録媒体においては、前記評価対象の文書は複数あり、前記プログラムは、前記各文書からタームを抽出し、抽出したタームで各文書を数値化して記憶しておき、以下の1)~2)の処理を前記複数の視点プロフィールについて実行して、前記評価値を求める。1)前記ある視点プロフィールの複数の関連ワードを数値化し、2)前記ターム数値化処理による値と前記視点プロフィール数値化処理による値に基づいて、当該視点プロフィールについての評価を演算する。したがって、前記文書データを客観的に、かつ正確に評価することができる。

【0023】15)本発明にかかるプログラムを記憶した記録媒体においては、前記タームによる数値化処理では、前記各文書を構成する語句について tfidf 値を求めて、重要語句を複数決定し、この複数の重要語句を用いて前記各文書の tfidf 値を演算して、前記重要語句の数と同じ次元の数値ベクトルが求められ、前記各視点プロファイルによる評価値演算処理では、前記各視点プロファイルにおける関連ワードの tfidf 値を演算して、前記関連ワードの数と同じ次元の数値ベクトルが各文書毎に求められ、前記両数値ベクトルの類似度が前記評価として演算される。したがって、前記文書データを客観的に、かつ正確に評価することができる。

【0024】16)本発明にかかる文書データ評価方法においては、視点プロファイルとこの視点プロファイルに関連する複数の関連ワードで構成される視点プロファイル対応情報を複数記憶しておき、前記文書データについて、前記各視点プロファイル毎の評価値を求めて、各視点プロファイルと同じ次元を持つ数値ベクトルを当該文書データの評価値とする。これにより、前記文書データを客観的に、かつ正確に評価することができる。

【0025】17)本発明にかかる文書データ評価装置においては、視点プロファイルとこの視点プロファイルに関連する複数の関連ワードで構成される視点プロファイル対応情報を複数記憶する対応情報記憶手段と、評価対象の文書データについて、前記各視点プロファイル毎の評価値を求めて、各視点プロファイルと同じ次元を持つ数値ベクトルを当該文書データの評価値として決定する評価手段を備えている。これにより、前記文書データを客観的に、かつ正確に評価することができる。

【0026】

【発明の実施の形態】1. 機能ブロック図の説明本発明の一実施形態を図面に基づいて説明する。図1に示す文書データ評価装置を含むコンピュータシステム1は、文書記憶手段3、対応情報記憶手段5、評価手段7、クラスタリング手段9、抽出手段11、相関関係演算手段13、推移演算手段15、予測手段17および表示手段19を備えている。

【0027】文書記憶手段3は、複数の文書が記憶される。対応情報記憶手段5は、視点プロファイルとこの視点プロファイルに関連する複数の関連ワードで構成される視点プロファイル対応情報を複数記憶する。評価手段7は、前記評価対象の各文書データについて、前記各視点プロファイル毎の評価値を求めて、各視点プロファイルの評価値を決定する。

【0028】本実施形態においては、評価手段7は以下のようにして、評価値を決定する。前記各文書からタームを抽出し、抽出したタームで各文書を数値化して記憶しておく。そして、以下の 1)～2)の処理を前記複数の視点プロファイルについて実行して、前記評価値を求める。1)前記視点プロファイルの複数の関連ワードを数値化し、2)前記ターム数値化処理による値と前記視点プロファイル数値化処理による値に基づいて、当該視点プロファイルについての評価を演算する。

【0029】クラスタリング手段9は、前記各文書データの評価値に基づいて、クラスタリングする。抽出手段11は、異なるクラスタに分類された文書データを比較して、あるクラスタに分類された文書データから、そのクラスタに関して特徴的なキーワードを抽出する。かかる抽出されたキーワードは表示手段19に表示される。

【0030】相関関係演算手段13は、前記各文書データの評価値に基づいて、前記複

数のプロファイルの相関関係を求める。かかる相関関係は表示手段19に表示される。推移演算手段15は、当該文書データを時系列に並べて、各視点プロファイルに関する前記各文書データの評価値の時間的変化を求める。かかる評価値の時間的変化は表示手段19に表示される。予測手段17は、前記求めた前記各文書データの評価値の時間的変化に基づいて、当該視点プロファイルにおける変化を予測する。かかる予測結果は表示手段19に表示される。

【0031】また、推移演算手段15によって演算された前記求めた各視点プロファイルに関する前記各文書データの評価値の時間的変化に基づいて、評価手段11は、変化の大きな時期の文書データを特定し、特定した文書データから、その文書に関して特徴的なキーワード、例えば、他の文書データには存在しないユニークなキーワードを抽出する。

【0032】本実施形態においては、抽出手段11、相関関係演算手段13、推移演算手段15、および予測手段17からのデータを報知する報知手段として表示手段19を用いたが、これ以外の報知手段を採用してもよい。

【0033】2. ハードウェア構成(2.1)概略図1に示すデータ処理装置1のハードウェア構成について説明する。図2に示すコンピュータシステム40は、入力装置41、制御装置43、表示装置45および記憶装置47を備えている。入力装置41は、各種の命令を入力するためのものである。記憶装置47には、与えられた命令に基づいて所定の処理を行うプログラムが記憶される。制御装置43は、記憶装置47に記憶されたプログラムに基づいて所定のデータ処理を行う。

【0034】(2.2)詳細図3に、図2に示すコンピュータシステム40をCPUを用いて実現したハードウェア構成の一例を示す。

【0035】コンピュータシステム40は、CPU23、メモリ27、ハードディスク26、CRT30、FDD25、キーボード28、マウス31およびバスライン29を備えている。CPU23は、ハードディスク26に記憶された制御プログラムにしたがいバスライン29を介して、各部を制御する。

【0036】この制御プログラムは、FDD25を介して、プログラムが記憶されたフレキシブルディスク25aから読み出されてハードディスク26にインストールされたものである。なお、フレキシブルディスク以外に、CD-ROM、ICカード等のプログラムを実体的に一体化したコンピュータ可読の記録媒体から、ハードディスクにインストールさせるようにしてもよい。さらに、通信回線を用いてダウンロードするようにしてもよい。

【0037】本実施形態においては、プログラムをフレキシブルディスクからハードディスク26にインストールさせることにより、フレキシブルディスクに記憶させたプログラムを間接的にコンピュータに実行させるようにしている。しかし、これに限定されることなく、フレキシブルディスクに記憶させたプログラムをFDD25から直接的に実行するようにしてもよい。なお、コンピュータによって、実行可能なプログラムとしては、そのままのインストールするだけで直接実行可能なものはもちろん、一旦他の形態等に変換が必要なもの(例えば、データ圧縮されているものを、解凍する等)、さらには、他のモジュール部分と組合して実行可能なものも含む。

【0038】ハードディスク26には、プログラム記憶部26a、対応表記憶部26b、文書記憶部26c、評価記憶部26dを有する。プログラム記憶部26aには、後述するプログラムが記憶されている。対応表記憶部26bには、図4に示すような複数の視点プロファ

イルが記憶されている。各視点プロファイルは対応する複数の関連キーワードが記憶されている。文書記憶部26cには評価対象の文書が複数記憶されている。本実施形態においては、各文書は、作成日付、その文書のタイトルおよび各文書の内容で構成されている。評価記憶部26dには評価結果が記憶される。メモリ27にはその他、各種の演算結果等が記憶される。

【0039】3. フローチャートつぎに、ハードディスク26のプログラム記憶部26aに記憶されているプログラムについて、図5、図6を用いて説明する。以下では、視点プロファイルとして「好景気」、「不景気」、「財政」を用いて、各文書进行评估する場合を、例として説明する。

【0040】操作者は対応表作成処理をおこなう(図5ステップS1)。まず、操作者は一覧表示命令を与える。これにより、図4に示す対応表がCRT30に一覧表示される。操作者は、これから評価する全視点プロファイルが存在するか否か判断する。もし、いずれかの視点プロファイルが存在しない場合には、視点プロファイル作成を行う。具体的には、視点プロファイルおよびこれに対応する関連キーワードを入力するようにすればよい。また、存在する視点プロファイルについても、関連キーワードを追加削除する必要があるかを判断し、追加削除するものがあれば同様に関連キーワードの追加削除を行う。このように、操作者の興味のある視点からの視点プロファイルを作成して、評価することにより、最終的に求められる評価として、操作者の望む視点からの評価値を得ることができる。

【0041】つぎに、CPU23は、評価処理を行う(ステップS3)。評価処理の詳細フローチャートを図6に示す。CPU23は、選択視点プロファイル番号iを初期化する(図6ステップS11)。CPU23は、視点プロファイルの選択があったか否か判断しており(ステップS13)、視点プロファイルの選択があると、その視点プロファイル番号を記憶する(ステップS15)。

【0042】CPU23は、検索開始命令が与えられたか否か判断しており(ステップS17)、検索開始命令が与えられなければ、ステップS13に戻り、視点プロファイルの選択があるか否か判断する。かかる処理により、必要な視点プロファイルが複数記憶される。具体的には、図4に示す視点プロファイル「好景気」、「不景気」および「財政」の3つの視点プロファイルが記憶される。

【0043】操作者は、必要な視点プロファイルを選択すると、CRT30に表示される検索開始ボタン(図示せず)をクリックする。

【0044】CPU23は、ステップS17にて検索開始命令が与えられると、各文書からタームを抽出してベクトル化する(ステップS19)。本実施形態においては、以下のようにして、ベクトル化した。まず、各文書について形態素解析を行い、文書中に出現する単語を抽出する。抽出した単語について、tfidf法を用いてそのタームが、重要タームであるか否かを判断する。tfidf法とは、情報検索におけるキーワード決定の手法であり、ある文書中におけるそのタームの出現頻度を示すtf(term frequency)および全文書中で当該タームがいかに少ない文書でしか現れないかを示すidf(inverse document frequency)を用いて、タームの重み付けをする手法である。

【0045】このようにして抽出した重要タームについて、各文書におけるtfidf値を求め、重要タームの数と同じ次元の数値ベクトルとしてベクトル化する。例えば、重要タームが100ある場合に100次元の数値ベクトルが得られる。なお、文書によっては決定さ

れた重要タームを含んでいない場合がある。この場合には、その文書のその次元の値は0(疎)となる。CPU23は、このようにして得られた数値ベクトルをメモリ27に記憶しておく。

【0046】つぎにCPU23は、i番目の視点プロファイルをベクトル化する(ステップS21)。本実施形態においては、i番目の視点プロファイルの全関連キーワードをハードディスク26から読み出して、視点プロファイル内の各関連キーワードのtfと全文書におけるidfから、前記各関連キーワード毎のtfidf値を求め、前記関連キーワードの数と同じ次元の数値ベクトルとしてベクトル化した。

【0047】つぎに、i番目の視点プロファイルに対する各文書の類似度を演算する(ステップS23)。かかる類似度は、ステップS19で求めた数値ベクトルとステップS21で求めた数値ベクトルの内積を演算することにより求めることができる。

【0048】つぎにCPU23は、全視点プロファイルについて類似度演算が終了したか否か判断する(ステップS25)。全視点プロファイルについて類似度演算が終了していないければ、選択視点プロファイル番号iをインクリメントし(ステップS27)、ステップS21以下の処理を行う。

【0049】つぎに、CPU23は、全視点プロファイルについて類似度0でない文書の番号を特定する(ステップS29)。CPU23は、ステップS29で特定した文書を一覧表示する(ステップS31)。図7に特定された文書および各視点プロファイルの類似度の一覧を示す。

【0050】図6の評価処理によって得られた類似度に基づいて、図5ステップS5の推移演算処理、ステップS7の予測処理等の、分析処理が行なわれる。なお、本実施形態においては、推移演算処理と予測処理の双方を実行したが、いずれかだけでもよく、また、後述する他の処理を行うようにしてもよい。

【0051】推移演算処理(図5ステップS5)について説明する。本実施形態においては、逆の方向性を有する1対の視点プロファイルとして「好景気」と「不景気」を含んでいる。したがって、操作者は両者の差分を演算する命令を与える。これにより、図8に示すような、各文書の視点プロファイル「好景気」「不景気」の値の差分を得ることができる。かかる差分の変化を時系列で並べて、図9に示すような折れ線グラフを生成する。これにより、記憶されている文書における景気の変化の様子を操作者に報知することができる。

【0052】つぎに、予測処理(図5ステップS7)について説明する。図9に示すグラフから図10に示すように、将来、視点「景気」がどのように変動するかを予測することができる。

【0053】CPU23は、かかるステップS5、7の処理を表示することにより、処理を終了する。

【0054】このような文書の内容を視点プロファイルによって客観的に評価することにより、種々の統計的処理を実行することができる。視点プロファイルに対する変化の要因を得たい場合には、つぎのようにすればよい。表示されたグラフについて、マウス31によって、図11に示すように変化の度合いの大きな時期として、領域61を特定する。これにより、CPU23は、かかる時期の日付を有する文書を特定して、文書のタイトルを表示する。操作者が内容を検証したい文書を特定すると、画面上に特定された文書が表示される。操作者は表示された文書を開き、その内容を見て、前記変化

の要因を判断することができる。

【0055】なお、このように特定された文書からつぎのようにして、前記要因を判断するための要因関連タームを抽出するようにしてもよい。かかる特定された文書のタームのうちtfidf値の高いタームを抽出する。抽出したタームのうち、前記視点プロファイルの関連キーワードに該当しないタームを、tfidf値の大きな順に並べて表示する。操作者はかかる表示されたタームから、要因関連タームを選択する。なお、表示されたタームについては、タームだけでなくその前後の文章を表示するようにしてもよい。これにより、操作者は、要因関連タームの選択がより容易となる。

【0056】また、上記のように複数の視点プロファイルにおける各文書の評価として、以下のような処理を行ってもよい。

【0057】各文書の分類法としては、例えば、あらかじめ分類基準がない場合には、クラスタ分析を行えばよい。そして分類後は、各集団の典型例を抽出することにより、その集団の意味を推定することもできる。典型例の抽出するには、例えば、集団の中心に一番近い文書を抽出すればよい。図12に、図8に示すデータに基づいて、視点プロファイル「好景気」、「不景気」の2つの視点プロファイルに注目して、2次元座標上にプロットして、3つの集団にクラスタリングした例を示す。このように、複数の視点プロファイルによって各文書を数値化することにより、操作者の視点に合致した数値化が可能となり、各文書を操作者の視点に合致して分類することができる。

【0058】また、特定の視点プロファイルに関するベクトル要素値に注目して、ベクトル要素値の範囲をいくつかに分割して、各範囲に属する文書をまとめるクラスタリングも可能である。例えば、前記2つの視点プロファイル「好景気」、「不景気」について、それぞれ、しきい値を決定し、各文書をクラスタリングする。例えば図7に示す例では、好景気、および不景気のしきい値を0.4とする。これにより、図13に示すような4つのグループに分類することができる。このように、各文書をクラスタリングしてから、文書分析することにより、より正確に文書分析することができる。

【0059】なお、1つの視点プロファイルについて2以上のしきい値を用いて、3つ以上に分割してもよく、さらに、3以上の視点プロファイルを用いて分割するようにしてもよい。

【0060】また、クラスタリングの手法としては、従来から用いられている手法が採用でき、例えば、上記のような非階層的クラスタリング分析だけでなく、階層的クラスタリング分析するようにしてもよい。

【0061】さらに、視点プロファイルに加えて、文書の作成された日時、著者、特定のタームを含むもの、含まないもの等の任意の分類が可能である。

【0062】このような分類処理により、ある分類グループの特徴的なタームを抽出し、比較検討することにより、新たな知識を発見することも可能となる。例えば、あるグループに属する文書にのみ出現する特徴的タームをそのグループの特徴タームとみなすことができる。

【0063】4. 他の実施形態上記のように複数の視点プロファイルにおける各文書の評価を行うことにより、以下のような統計的処理をすることができる。

【0064】図7に示すデータから、各視点プロファイルが有する意味概念間の相関関係を求めることができる。例えば、この場合、「好景気」と「不景気」の相関係数は-0.30で、弱い負の相関関係があることが分かる。このように、各文書を視点プロファイ

ルで数値化することにより、視点プロファイル間の相関を知ることができる。

【0065】また、特定の視点に注目して、その視点に影響を与える他の視点を求めるためには、それぞれの視点プロファイルに対応するベクトル要素値を用いて重回帰分析を行い、標準回帰係数によって判断することができる。例えば、図14のように、視点プロファイル「財政」を目的変数に、視点プロファイル「好景気」と「不景気」を説明関数として重回帰分析を行うことにより、図14に示すような回帰係数を得ることができる。

【0066】かかる回帰係数から視点プロファイル「財政」に対しては視点プロファイル「不景気」の影響が大きいことがわかる。このように、各視点プロファイルに対する要因を数値で客観的に評価分析する事ができる。

【0067】さらに、文書をその日付けで分類して、各時間区分毎に標準回帰係数の推移を見れば、視点プロファイル間の影響の度合いの時間的变化を把握することができる。

【0068】本実施形態においては、前記得られた多次元ベクトルを評価値として採用したが、いくつかの視点プロファイルに対応するベクトル要素値の重み付き和を求めて、各文書の評価値としてもよい。

【0069】また、上記実施形態においては、各文書を時系列順に並べて、各文書の得点の推移を求めるようにしたが、さらに、複数の視点プロファイルに関する得点を演算して、それぞれの得点を各軸に割り当ててトレンドの多次元表示を行ってもよい。あるいは、他の数値データと組み合わせた多次元表示でもよい。この場合、プロットされた曲線は時間に関する陰関数となる。例えば、図10に示すグラフを縦軸に「株価」の割り当てることにより、株価と景気との関係进行を判断することができる。例えば、図15に示すようなデータが得られた場合、かかる曲線はおおよそ反時計回りに回っていることが分かるので、株価の将来を予測することもできる。

【0070】図15は時間を媒介変数とした場合であるが、時間に限られることなく、例えば人や物であってもよい。図16に、商品の販売個数と顧客からの商品毎の評価レポートの分析結果を、商品を媒介変数としてプロットした場合を示す。評価レポートの分析には商品进行评估するいろいろな視点プロファイルを使用することができ、この例では、デザインや機能性に関するプロファイルを作成し、それぞれの得点の総和を総合得点としてプロットしている。商品毎の販売個数データ71と、商品毎の総合得点72との関係から、商品を媒介変数にして、総合得点と販売個数との関係74を最終的に得ることができる。

【0071】上記実施形態においては、前記複数の視点プロファイルとして、逆の方向性を有する1対の視点プロファイルを採用したが、これに限定されない。

【0072】また、本実施形態においては、前記視点プロファイルとして状態を表すプロファイルとして「好景気」、「不景気」を用いたが、変化を表すプロファイルとして「増産」、「増益」等を用いてもよい。このような状態または変化を表すプロファイルを用いることにより、マイニングがより容易となる。なお、視点プロファイルについては、操作者が望む視点プロファイルであれば、これらに限定されず、どのようなものであってもよい。

【0073】なお、本実施形態においては、全視点プロファイルについて、類似度が0でない文書を抽出したが、全視点プロファイルのうち少なくとも1つ以上について類似度0でない文書番号を特定するようにしてもよい。

【0074】なお、各文書からタームを抽出してベクトル化する処理については、全文書

の内容を確定できれば、あらかじめ処理することも可能であるので、新たな文書が追加された時に実行して、記憶しておいてもよい。

【0075】また、本実施形態においては、各文書から重要タームを抽出するようにしたが、タームであればこれに限定されず、例えば重要タームだけでなく、全タームを抽出するようにしたり、ある抽出基準で抽出するようにしてもよい。

【0076】また、本実施形態においては、他の文書データには存在しないユニークなキーワード等を抽出することより、あるクラスタに関して特徴的なキーワードを抽出するようにしたが、例えば、全クラスタに共通に出現するキーワードや、特定のクラスタにだけ大きな値を有するキーワードを抽出するようにしてもよい。

【0077】本実施形態においては、2つの数値ベクトルの類似度を、両数値ベクトルの内積を演算することにより決定したが、両数値ベクトルのコサイン値を類似度としてもよい。

【0078】また、本実施形態においては、日本語の文書の場合について説明したが、他の言語、例えば、英語、中国語、韓国語等についても同様に適用することができる。

【0079】本実施形態においては、図1に示す機能を実現する為に、CPU23を用い、ソフトウェアによってこれを実現している。しかし、その一部もしくは全てを、ロジック回路等のハードウェアによって実現してもよい。

【0080】このように、文書を操作者の興味がある複数の視点プロファイルに関する得点によって数値ベクトル化している。これにより、タームの出現頻度に依存した値で表す場合と比べて、文書内容を操作者が容易に把握することができる。また、数値ベクトル化された文書を各種の統計的解析手法を用いて分析することができる。さらに、分析結果の意味付けが容易となる。

図の説明

【図面の簡単な説明】

【図1】本発明にかかる文書データ評価装置1の機能ブロック図である。

【図2】図1に示す文書データ評価装置のハードウェア構成の一例を示す図である。

【図3】図2に示す文書データ評価装置40をCPU23を用いて実現したハードウェア構成の一例を示す図である。

【図4】視点プロファイルと関連キーワードの対応を示す図である。

【図5】文書評価処理の全体フローチャートである。

【図6】評価処理の詳細フローチャートである。

【図7】評価結果を示す図である。

【図8】評価結果を示す図である。

【図9】評価値の時系列変化を示す図である。

【図10】評価値の時系列変化を示す図である。

【図11】文書を取り出す時期を特定する状態を示す図である。

【図12】クラスタリングの一例を示す。

【図13】クラスタリングの一例を示す。

【図14】重回帰解析結果を示す。

【図15】景気と株価との関連を示す。

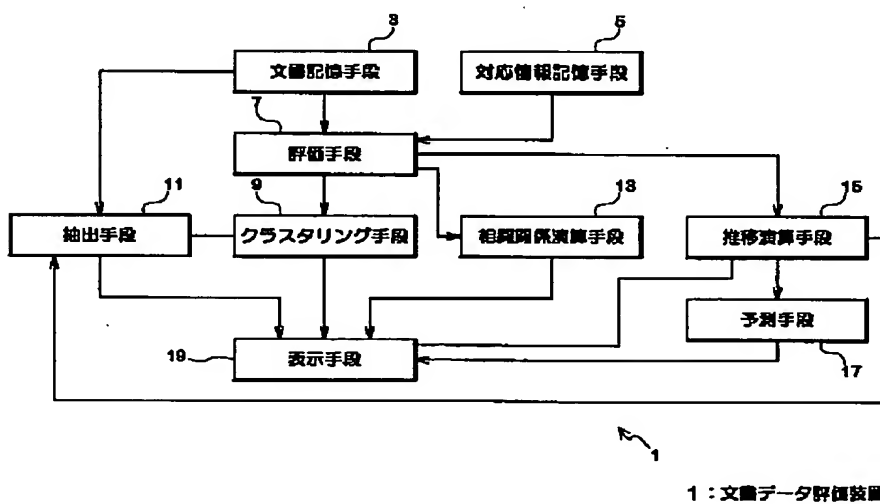
【図16】商品の評価と販売個数との関係を示す。

【符号の説明】

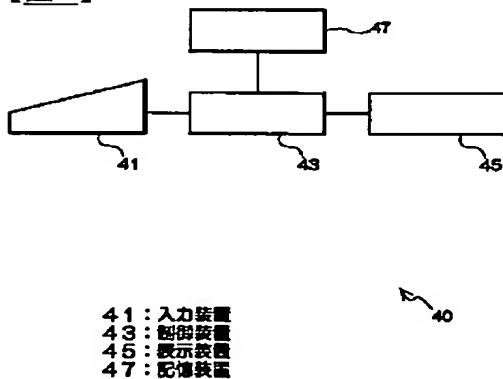
- 3.....文書記憶手段
- 5.....対応情報記憶手段
- 7.....評価手段
- 9.....クラスタリング手段
- 11.....抽出手段
- 13.....相関関係演算手段
- 15.....推移演算手段
- 17.....予測手段
- 19.....表示手段
- 23...CPU
- 27...メモリ

図面

【図1】



【図2】

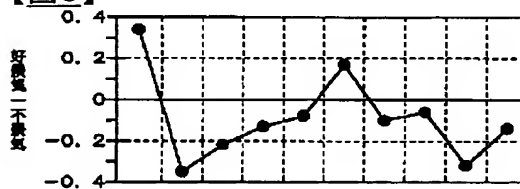


【図4】

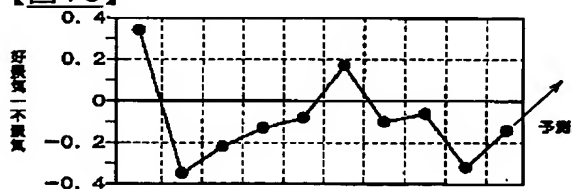
<対照表>

| No | 視点プロフィール | 関連キーワード |
|----|----------|-----------------------------|
| 1 | 好景気 | 好調、堅調、増産、増収、増益 |
| 2 | 不景気 | 不調、不振、減産、減収、減益、落ち込み、冷え込み |
| 3 | 財政 | 財政、歳入、歳出、収収、財政投融资、交付税、財源、国債 |
| 4 | スポーツ | 野球、テニス、水泳、サッカー、バレー |
| 5 | 政治 | 政府、行政、大臣、国会、永田町、省庁、選挙 |

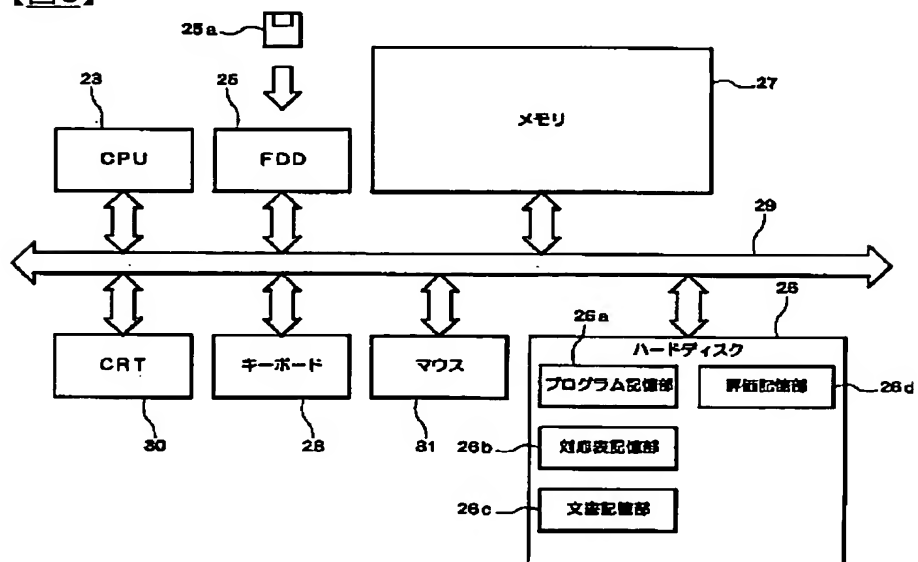
【図9】



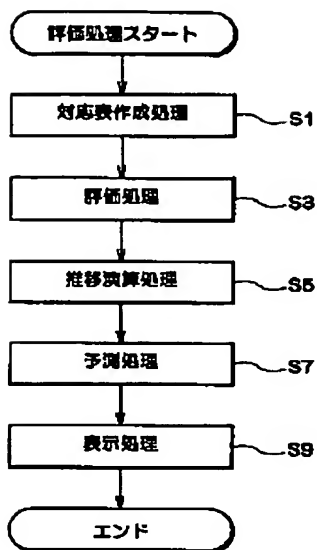
【図10】



【図3】



【図5】



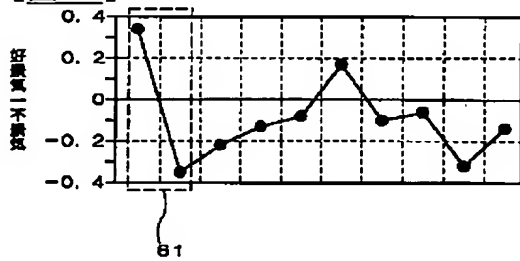
【図7】

| 文書 | | | 類似度 | | |
|----|---------|---------|------|------|------|
| No | 日付 | タイトル | 好景気 | 不景気 | 財政 |
| 1 | 98/1/1 | カルガリー大会 | 0.55 | 0.21 | 0.44 |
| 2 | 98/2/11 | 半導体 | 0.40 | 0.75 | 0.85 |
| 3 | 98/3/1 | 鳥獣の恐怖 | 0.39 | 0.61 | 0.51 |
| 4 | 98/4/1 | 家庭用ゲーム機 | 0.35 | 0.48 | 0.44 |
| 5 | 98/5/1 | 大型側産 | 0.32 | 0.39 | 0.38 |
| 6 | 98/6/1 | 旅行業界 | 0.42 | 0.25 | 0.21 |
| 7 | 98/7/1 | ヤオハン | 0.29 | 0.39 | 0.24 |
| 8 | 98/8/1 | 大学内紛 | 0.28 | 0.34 | 0.34 |
| 9 | 98/9/1 | 米価暴落 | 0.27 | 0.59 | 0.35 |
| 10 | 98/10/1 | 格闘 | 0.31 | 0.46 | 0.67 |

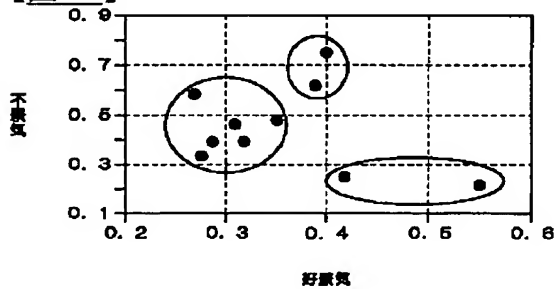
【図8】

| No | 日付 | タイトル | 好感度 | 不快感 | 好感度-不快感 |
|----|---------|---------|------|------|---------|
| 1 | 98/1/1 | カルガリー大会 | 0.55 | 0.21 | 0.34 |
| 2 | 98/2/11 | 半導体 | 0.4 | 0.75 | -0.35 |
| 3 | 98/3/1 | 鳥飼の招布 | 0.39 | 0.61 | -0.22 |
| 4 | 98/4/1 | 家庭用ゲーム機 | 0.35 | 0.48 | -0.13 |
| 5 | 98/5/1 | 大塚樹彦 | 0.32 | 0.39 | -0.07 |
| 6 | 98/6/1 | 旅行業界 | 0.42 | 0.25 | 0.17 |
| 7 | 98/7/1 | ヤオハン | 0.29 | 0.39 | -0.1 |
| 8 | 98/8/1 | 大学内紛 | 0.28 | 0.34 | -0.06 |
| 9 | 98/9/1 | 米田子篇 | 0.27 | 0.59 | -0.32 |
| 10 | 98/10/1 | 稲旗 | 0.31 | 0.46 | -0.15 |

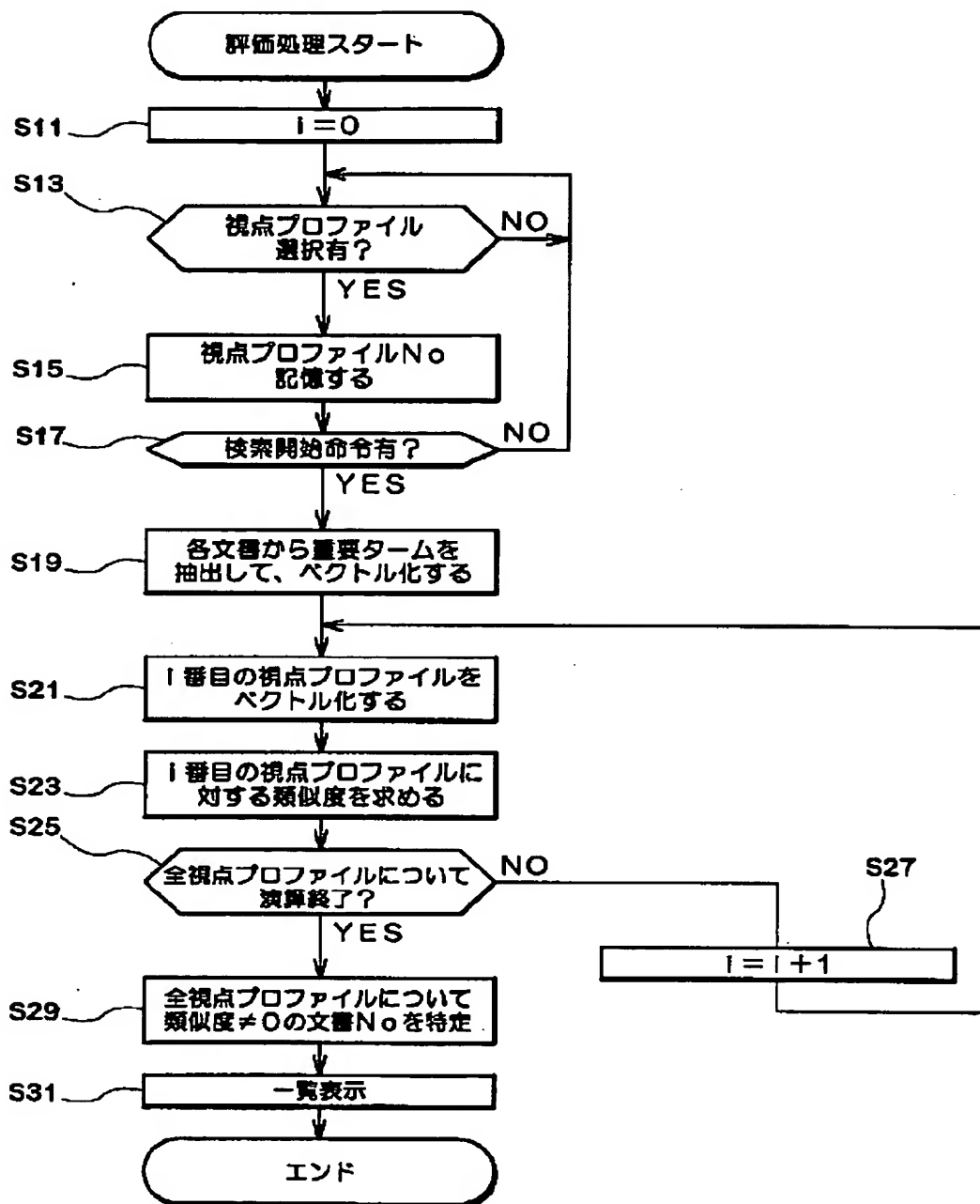
【図11】



【図12】



【図6】



【図13】

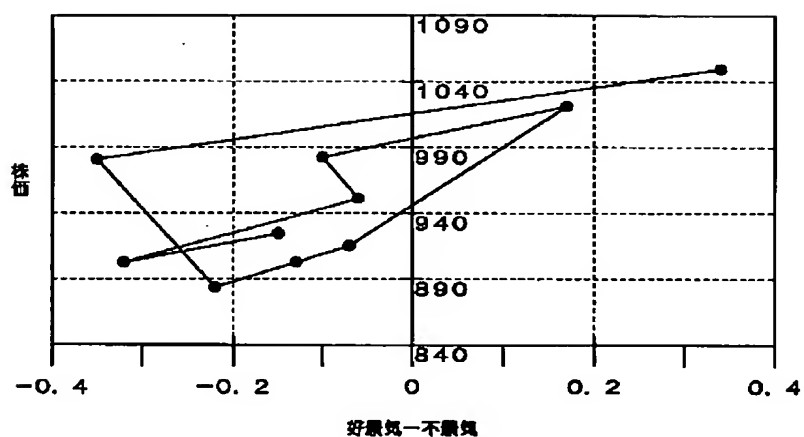
| | 不景気 0.4未満 | 不景気 0.4以上 |
|--------------|-------------------|---------------------------|
| 好景気 0.4未満 | 文書5 文書7 文書8 | 文書3 文書4 文書9 文書10 |
| 好景気 0.4以上 | 文書1 文書6 | 文書2 |

【図14】

| | 固有係数 | 標準固有係数 |
|-------|--------|--------|
| 好景気 | 0.936 | 0.411 |
| 不景気 | 0.921 | 0.794 |
| (定数項) | -0.304 | |

【図15】

景気状態と株価のプロット



【図16】

